NOISE REDUCTION FOR FACE IDENTIFICATION IN VIDEOS

by

NEGAR HASSANPOUR

B.Sc. Electrical (Control) Engineering
University of Tehran, Tehran, Iran.

THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
MATHEMATICAL, COMPUTER, AND PHYSICAL SCIENCES
(COMPUTER SCIENCE)

UNIVERSITY OF NORTHERN BRITISH COLUMBIA

2015

Abstract

The wealth of information extracted from a sequence of frames in a video provides samples of the subject in different illuminations, head poses, and facial expressions. However, various sources can impose noise on data (e.g., occlusion, low resolution, and face detection failure).

In this thesis, a novel framework is proposed that employs the well studied concepts in quantum probability theory to design a representation structure capable of making inferences with multiple sources of uncertainty. The dual extension of this framework is aimed at reducing the effect of noisy frames in a video. It is also used to guide the sampling process in a novel learning scheme, called specialization–generalization, which is designed to support efficient learning, as well as neutralizing the effect of noisy samples in the identification process.

The contributions of this thesis are not method-specific and can be utilized for enhancement of other face identification approaches in the literature.

Contents

List of Figures

List of Figures

Acknowledgements

First, I would like to thank my supervisor, Professor Liang Chen for his support, encouragement, patience, and help in both my education and life. His insight and guidance was what made this thesis possible, and I am grateful to him for all the academic achievements I have had during this degree. I have learned an enormous amount from working with him in terms of how to be both a researcher and a supervisor. I hope to follow his example in the future.

A number of sources funded my research. In particular, I would like to thank the University of Northern British Columbia, Mitacs, and NSERC for the financial support that they provided either directly or through grants awarded to Dr. Chen.

Thanks to all my friends in Prince George, Nahid and Mani, Mona and Rahim, Mojtaba, Dhawal, and Behrooz. You were what made living here so amazingly rewarding.

I am deeply grateful to my family. My father, Nemat Hassanpour, who emphasized the importance of education, and who instilled in me the inspiration to set high goals and the confidence to achieve them. My mother, Soheyla Norouzi, who has always been a source of motivation and strength, and my role-model for hard work, persistence, and personal sacrifices. And my sister, Bahar Hassanpour, who has been my emotional anchor through my life.

Finally, I would like thank my lovely husband, Samad Kardan. It was difficult for

us since we lived across the province from one another for most of the time of this degree. Samad supported me at all times in a loving, empathetic, optimistic, calm, and patient manner. This thesis is dedicated to him.

Chapter 1

Introduction

Face identification is considered as one of the most important applications of image analysis and understanding and has received a lot of attention from the machine vision community through the past several years. The classical face identification task involves identifying a subject from a single image and with only a few training samples available. For such configuration, these sample images must be carefully recorded in a controlled environment. However, in real-world applications, such good quality samples are not easily attainable.

Fortunately, with the extensive availability of digital imaging devices, sufficient data is accessible to allow the recognition process to be based on image-set to image set matching. Image-set could be either a collection of single shot images featuring a person, or a sequence of frames in a video.

The wealth of information extracted from a sequence of frames in a video featuring an individual's face offers the potential to overcome the recognition errors that may

occur due to the imperfect quality of the images, which is an absent privilege in the single shot face identification tasks. In this sense, image-set based face identification in general and video based face identification in particular is potentially more promising than using single-shot images. This type of face identification tends to be more robust since the recognizer gets to see many more possible variations in appearance of the subject.

However, with image-sets,a new challenge emerges due to the uncertainty on how well each image represents the individual who it is associated with. Uncertainties may be due to (i) poor quality of the image (e.g. low resolution, illumination, contrast, etc.), (ii) partial existence of the face in the image's field of view (e.g. occlusion, pose, etc.), and (iii) failure of the face detector algorithm to accurately spot the face. All in all, uncertainties are imposed due to the fact that each image may not fully characterize the subject's face.

Therefore, it is important to devise algorithms that can fully and efficiently exploit the available data. This can be done through reducing the effect of noisy samples by designing a representation structure capable of relaxing the noise in each sequence, complemented by developing a recognition procedure that rejects the wrong decisions affected by noise. Devising a principled way to systematically deal with the uncertainties on how well each frame can represent the individual has not yet been directly addressed in the literature. This is the focus of this thesis.

1.1 Overview of this thesis

The quantum probability theory offers a powerful framework for representing information and making inferences in systems with multiple sources of uncertainty. To deal with uncertainties, quantum theory proposes to consider an ensemble of all possible initial states simultaneously as it attempts to find the most probable outcome (i.e., event) of the system. To achieve this, the mathematical formalism of quantum theory extends the ordinary logic by the concept of *simultaneous decidability* – a concept introduced by von Neumann and Beyer [1955] – which allows physicists to continue reasoning while considering the uncertainty associated with the state of the system. Inspired by this line of research, we propose a Quantum Probability Inspired Framework (referred to as **QPIF**) for face identification in videos which uses the quantum probabilities to address the underlying uncertainties associated with the representativeness of each image.

It is suggested that the mathematical foundations of quantum probabilities would construct a sound knowledge representation system. In this representation, information extracted from the images of the known identities form subspaces in a Hilbert space, where each subspace represents one subject of known identity. When presented with an image-set that belongs to a subject of unknown identity, an ensemble of uncertain states is generated from the image-set. Recognition is posed as an optimization problem to find the best-ranked subspace to which the ensemble of states corresponds.

Next, we will provide a dual extension of the quantum framework called Ensemble of Abstract Sequence Representatives (referred to as **EASR** – nomenclature will be described in section 4.2). In EASR approach, the representation structure of all images – either of known or unknown identity – is the same as the concept of initial state in

the quantum formalism. This approach towards data representation would uniformly relax the noise in all of the raw data points by transferring them into a higher level representation space.

Each EASR is built through a process that includes sampling and superposition of the raw images in order to reduce noise. This is followed by a filtering mechanism to deal with outliers. Similar to the majority of the image-set based face identification approaches (e.g., works by Wang and Chen [2009], Wang et al. [2012b], Kim et al. [2007], Hu et al. [2012], and Yang et al. [2013]) that use a single structure to model each image-set, each EASR tries to model the variations in appearance of the subject in an image-set. Similarity of EASRs is calculated as the distance between each pair of train (of known identity) and test (of unknown identity) image-sets. Identification is performed by finding the most similar representation of a known candidate to different representations of the unknown subject, and then aggregating the identification results of all candidates via majority voting over the different representations generated for the unknown subject.

Although EASRs reduce the noise in data, they are linear representations and are not capable of capturing the underlying non-linear structure of the data. Therefore, on top of the EASR representation method, we introduce an ensemble of binary Gaussian process (GP) models in a one-versus-rest setting for capturing the underlying non-linearity in the data. To reduce the amount of noise presented to the GP models during the training, we use a learning scheme called specialization – generalization. The specialization step attempts to find a subset of training data samples such that the highest discriminative power is achieved (i.e., only select the most challenging samples to train to the classifier, not all the training samples). The generalization step attempts to reduce the effect of possibly noisy training samples by re-training

the model on those unseen samples from training set that the model failed to classify correctly, thus, making sure that the model generalizes well. Finally, a fast identification process combines predictions of both methods to identify the subject in the probe sequence. We would refer to this hybrid identification system as **EASR+GP**.

1.2 Main contributions

The main contribution of this work is two-fold:

First, we propose two representation structures for image-sets that are designed to minimize the effect of noisy frames. Inspired by the concepts in quantum probability theory, QPIF and its dual extension EASR are designed such that the impact of those frames that are not useful for the identification task (probably due to occlusion, low resolution, or failure of the face tracker algorithm) is reduced.

Second, a novel learning scheme was proposed for efficient training of an ensemble of binary Gaussian process models. This learning scheme selectively samples from the training data in order to not only increase the discrimination power of the classifier, but also to build the models using the least possible computational cost and with minimum introduction of noise.

Assessment of the proposed method on three publicly available benchmark datasets demonstrates significantly higher performance compared to the previous methods in the literature including state-of-the-art.

Part of the contents of this thesis has been published in the 11th IEEE International Conference on Automatic Face and Gesture Recognition (Hassanpour and Chen

[2015a]), and another part is published as a technical report at the Computational Intelligence Laboratory – University of Northern British Columbia (Hassanpour and Chen [2015b]).

1.3 Organization of this thesis

The rest of this document is organized as follows:

In chapter 2 the previous works on the three main aspects of this research are discussed. This includes Image-set based Face Identification, Quantum Theory in Information Retrieval, and Gaussian Processes in Computer Vision.

In chapter 3, the background information on which the foundation of the proposed methods is based on is explained. In the first two sections of this chapter, the mathematical concepts of Quantum Theory and Gaussian Processes are discussed. In the the third section, the Face Detection algorithms that are used to locate faces in a video frame are briefly introduced. The fourth section touches on the Feature Extraction methods. And in the final section, a brief introduction on the Welch t-test is provided. This statistical test is employed to check whether the improvements are significant.

In chapter 4, the proposed algorithms are elaborated in light of the mathematical foundations discussed in chapter 3. The proposed methods are: (i) Quantum Probability Inspired Framework (QPIF), (ii) Ensemble of Abstract Sequence Representatives (EASR), and (iii) Ensemble of Gaussian Process Models on top of the EASR approach (EASR+GP) with a hierarchical approach on the identification process.

In chapter 5, the experimental setup for performance evaluation of the proposed methods against the previous methods in the literature is described. This chapter includes an introduction of the Datasets (namely, Honda/UCSD, CMU-MoBo, and YouTube Celebrities), as well as the Evaluation Settings that were equally set for evaluating all methods to allow for fair comparison. These settings include: face detection algorithm, image resolution, extracted features, and partitioning the available data into train and test subsets.

In chapter 6 the identification accuracies of the three proposed approaches as well as the most successful methods in the literature are reported. Additionally, the average computation times of all methods are noted in order to give a sense of each algorithm's computational complexity.

Finally, in chapter 7 this document is concluded by summarizing the main contributions of the proposed methods and highlighting the future directions of this research.

Chapter 2

Previous Work

In this chapter, the previous works conducted in the literature are discussed. In the first section, an overview of the current image-set based face identification techniques is provided. Next, in the second section, a brief survey summarizing the applications of quantum probability theory in computer science in general and machine learning tasks in particular is provided. Finally, the third section gives a brief history on employing Gaussian process models to solve machine vision problems.

2.1 Image-set based Face Identification

In most of the published studies, the task of image-set based face identification is addressed in two steps: (i) representation of the image-sets, (ii) finding a suitable similarity measure between them. In the following of this section, the basic concepts of the most known approaches (including the state of the art) are elaborated

2.1.1 Representation

It is essential to be able to represent every image-set (often with varying number of images inside) in a unified manner, so that different image-sets can be compared with each other. The aim of a representation structure is to provide a well-defined method to transfer information embedded in a set of images into a unified structure and preserve as much information as possible. Representation structure of the image-sets can be either parametric or non-parametric.

Parametric methods

Parametric methods attempt to represent each image-set with a data-driven distribution function. For instance, Arandjelovic et al. [2005] fit a Gaussian mixture model to each image-set and use this distribution as the representative of the respective image-set. Parametric methods, however, suffer from the assumption that all image-sets representing the same identity are drawn from the same distribution. However, empirical results have shown that this is most likely not the case. Therefore, majority of the current works design a non-parametric representation structure.

Non-parametric methods

The non-parametric representations are divided into two categories: linear and non-linear.

Linear Representations. Most notable linear methods include:

- Mutual Subspace Method *MSM*, proposed by Yamaguchi et al. [1998], constructs a linear subspace for each image-set and calculates the similarity using the Euclidean angle between the two subspaces.

- Discriminant Canonical Correlations *DCC*, proposed by Kim et al. [2007], finds an optimal discriminant function that transforms the image-sets into another space in which the within-class canonical correlations are maximized while the between-class canonical correlations are minimized.

Non-linear Representations. Non-linear methods include:

- Constrained Mutual Subspace Method *CMSM*, proposed by Fukui and Yamaguchi [2005], constructs a constrained subspace that only includes the effective components of the input images for recognition (using principal component analysis), and measures the similarity between image-sets as the multiple canonical angle between the two subspaces.

- Kernel Grassmannian Distance *KGD*, proposed by Wang and Shi [2009], is a kernel generalization of the Grassmannian distance in order to capture the non-linear structures in the image-sets.

- Manifold Discriminant Analysis *MDA*, proposed by Wang and Chen [2009], forms the subspaces for each image-set with locally linear models (i.e., manifolds) and attempts to learn an embedding space, where each manifold is compact but manifolds of different classes are as separated as possible.

- Manifold-Manifold Distance *MMD*, proposed by Wang et al. [2012b], formulates the recognition task as computation of distance between two locally linear subspaces of data (i.e., manifolds).

2.1.2 Similarity Measurement

Once every image-set is represented in a unified manner, we can start to find out which image-sets are the most similar to each other, and consequently, develop a way to perform tasks such as identification. The method of similarity measurement depends on the representation method. If the representation is parametric, as in the work by Arandjelovic et al. [2005], in which each image-set is represented as a Gaussian mixture model, the similarity between a pair of image-sets is measured by calculating the between-set distribution distance (e.g., Kullback-Leibler divergence).

Measurement of similarity in non-parametric representations can be divided into three categories:

Examplar Based. The first category of similarity measurement methods are the ones that are based on calculating the distance between representatives of the two image-sets. Instances include:

- Affine/Convex Hull based Image-set Distance *AHISD* and *CHISD*, proposed by Cevikalp and Triggs [2010], represents each image-set by an affine/convex hull derived by spanning the subspace using the images in the set. The similarity is measured as the distance between the closest exemplars of each image-set.

Whole Structure. Similarity can also be measured based on the representation structure as a whole. Instances include:

- Covariance Discriminant Learning *CDL*, proposed by Wang et al. [2012a], represents the image-set by its covariance matrix (i.e., second-order statistic). This

formulates the problem as classification in the Riemannian manifold. A proposed function converts the covariance matrix from Riemannian manifold to Euclidean space where measurement of similarity is straightforward.

Both Examplar and Structure.

- Sparse Approximated Nearest Point *SANP*, proposed by Hu et al. [2012], proposes a similarity measurement method that utilizes both the structural information of the image-sets, as well as their representatives. The kernel extension of this approach, *KSANP*, allows for modelling the complex non-linear structures that are embedded in the data.

- Regularized Nearest Points *RNP*, proposed by Yang et al. [2013], models each image-set as a regularized affine hull and measures the similarity between the two sets by calculating the distance between the nearest points between the two hulls representing each image-set.
  RNP is an improvement over *SANP* in terms of complexity reduction.

2.1.3 Other Methods

Some recent methods have a holistic approach towards their representation structure which is complemented by their own representation-specific approach for similarity measurement. Instances include:

- Dictionary-based face identification from Video *DFRV*, proposed by Chen et al. [2012], captures variations in videos and records them in a dictionary while removing their redundancies. Identification is performed via majority voting.

- Mean Sequence Sparse Representation-based Classification *MSSRC*, proposed by Ortiz et al. [2013], performs a joint optimization to determine a linear relationship between all available training images to build its model.

- Joint Sparse Representation *JSR*, proposed by Cui et al. [2014], represents all the frames in a probe video sequence as an ensemble to suppress the effect of noise for a more stable recovery.

- Image-Set based Collaborative Representation and Classification *ISCRC*, proposed by Zhu et al. [2014], models the probe image-set as a convex or regularized hull and calculates the distance to the image-sets in the gallery considering the correlation between these two.

2.2 Quantum Theory in Information Retrieval

In the past two decades, quantum theory has found its way through theoretical computer science problems. From algorithms for database search (work by Grover [1996]) to decision theory (work by Pothos and Busemeyer [2009]), game theory (work by Piotrowski and Sładkowski [2003]), and information retrieval (work by Piwowarski et al. [2010]) to name a few.

The quantum information retrieval framework, for instance, focuses on representing queries and documents in terms of quantum probability theory in order to deal with the uncertainty imposed by ambiguous queries. Ambiguous queries might include cases such as polysemy and/or partial expression of the information need (consult work by Piwowarski et al. [2010] for more details).

2.3 Gaussian Processes in Computer Vision

Gaussian Process (GP) models have been previously used in several machine vision related applications, including:

- Human pose estimation: given an image, estimate the 3D location and orientation of the body parts (see work by Ek et al. [2008])
- Flow estimation: model the trajectory of a moving object (see work by Kim et al. [2011])
- Object recognition in an active learning paradigm: Kapoor et al. [2007] use GP confidence estimates at unlabelled data points in an active learning paradigm for interactive labelling. This active learning approach is of interest for datasets in which abundant unlabelled data is available, given that manual labelling is often expensive and/or time consuming in large datasets.

To the best of our knowledge, Gaussian process models have not yet been used by the machine vision community to address the task of video-based face identification.

Chapter 3

Background Information

3.1 Mathematics of Quantum Theory

In physics terms, quantum theory provides a mathematical description of matter at atomic and subatomic length scales, where there is uncertainty about the state of the particles. The reason is that at such length scales, the measurements cannot provide a rigid knowledge about the initial state of the system. To deal with such uncertainty, quantum theory proposes to consider an ensemble of all possible initial states simultaneously. To achieve this, the mathematical formalism of quantum theory extends the ordinary logic by the concept of *simultaneous decidability* – a concept introduced by von Neumann and Beyer [1955] – which allows physicists to continue reasoning while considering uncertainty about the state of the system.

Quantum probability is stated to be a geometrical extension of the classical probability theory. In classical probability theory, the probability space of a physical system

corresponds to a categorical property of the system. It is defined as a **discrete** set of all possible events (i.e., outcomes) that may occur if the system is in each known initial state. Quantum probability theory, on the other hand, employs the Hilbert space H, that is, a vector space together with an inner product as the probability space. Each probabilistic event is defined as a *continuous* subspace S in H.

Any theory consists of a set of primitive concepts and the relations between them which must be mapped to the real world of experience. According to Ballentine [1970], the primitive concepts of quantum theory are: (i) *"state vector"* which represents the initial state of the system, and (ii) *"event subspace"* which represents the probabilistic event that may occur to the system based on its initial state.

In the following, we will elaborate on these two concepts.

3.1.1 Initial State

Each state of a quantum system is defined as a unit feature vector ϕ in an n dimensional Hilbert space. The vector ϕ represents a *pure* state of the system. The initial state of the system may be a result of the superposition of N *pure* states, i.e., *mixed* state, following (3.1). The *mixed* states share the same properties as *pure* states: both are n dimensional and of unit length.

$$\psi = \frac{\sigma}{\|\sigma\|} \quad , \quad \sigma = \sum_{i=1}^{N} \phi_i \tag{3.1}$$

In case the initial state is not unique[1], an ensemble of all possible state vectors

[1]The cat in Schroedinger's paradox is both dead and alive concurrently, unless a deterministic

ψ is considered to contribute in derivation of the outcome of the system. Each ψ_i is associated with $p(\psi_i)$ that determines the probability of ψ_i being the true state. If the states are mutually exclusive[2], the probability distribution over the ensemble can be represented as a diagonal matrix $p(\psi)$:

$$p(\psi) = \begin{pmatrix} p(\psi_1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & p(\psi_M) \end{pmatrix}, \quad tr(p(\psi)) = 1 \tag{3.2}$$

where M is the number of all possible initial states. The ensemble of states in quantum formalism is defined by aggregation of the participating state vectors and their probabilities in form of a *density operator*:

$$\rho = \sum_{i=1}^{M} p(\psi_i)\psi_i\psi_i^T \tag{3.3}$$

Hence, state of a quantum system can be fully represented by an ensemble of all possible states. Ballentine [1970] notes that the concept of ensemble would allow for continuation of reasoning in uncertain environment and it is known as the *statistical interpretation* of the quantum theory.

---

measurement is done.

[2]In case the system cannot be in two or more states at the same time, all possible initial states of the system are mutually exclusive. Otherwise, $p(\psi)$ is not diagonal and the ensemble of initial states is considered to be *entangled* – out of the scope of this study.

3.1.2 Probabilistic Event

The outcome of a physical system is defined as a probabilistic event. In terms of quantum probability, it is represented as a subspace S in H. Probability of the event S given that the system is in initial state ψ_i is derived as the projection of vector ψ_i onto the subspace S:

$$q(S|\psi_i) = \|\hat{S}\psi_i\|^2 = \psi_i^T \hat{S}\psi_i = tr(\psi_i^T \hat{S}\psi_i) \tag{3.4}$$

where $\hat{S}$ is the projection matrix that gives a vector space projection from H onto subspace S.

3.1.3 Prediction

Prediction of the final event that may occur in the system is made by aggregating the effect of all possible states. In order to calculate the probability of event S happening, we need to marginalize (3.4) with respect to ψ:

$$\begin{aligned} q(S) &= \sum_{i=1}^{M} p(\psi_i) q(S|\psi_i) = \sum_{i=1}^{M} p(\psi_i) tr(\psi_i^T \hat{S}\psi_i) \\ &= \sum_{i=1}^{M} p(\psi_i) tr(\hat{S}\psi_i\psi_i^T) = \sum_{i=1}^{M} tr(\hat{S}p(\psi_i)\psi_i\psi_i^T) \\ &= tr(\hat{S}\sum_{i=1}^{M} p(\psi_i)\psi_i\psi_i^T) = tr(\hat{S}\rho) \end{aligned} \tag{3.5}$$

The above equation suggests that the probability of an event in quantum theory is derived as a weighted sum of the projections of all possible states over the event

subspace. This probability is calculated for all possible candidate events ($S_i, i = 1 : M$) and the event with the highest $q(S_i)$ would be reported as the prediction of the model on which event would occur.

3.2 Gaussian Processes

A Gaussian process is a generalization of the Gaussian probability distribution and is a Bayesian alternative to the kernel methods such as Support Vector Machines. Since models learned by GP are non-parametric, any hard assumptions on the structure of the model are safely avoided (e.g. assuming all data points are drawn from the same distribution, i.e., parametric models described in chapter 2). In this section, we briefly discuss GPs for regression and classification following the notation used by Rasmussen and Williams [2006] and Murphy [2012].

3.2.1 Regression

In supervised learning, regression attempts to predict the continuous quantities based on a set of observations. Formally, given a set of samples $X = \{x_1, x_2, ..., x_N\}$, where each x_i represents a feature vector, and the output of the unknown function at those points $y = \{y_1, y_2, ..., y_N\}$, we are interested to find the output of this unknown function at X_* data points.

The Gaussian process solution for regression assumes that a latent function $f(x)$ exists such that $y = f(x) + \epsilon$, where $\epsilon \sim (0, \sigma_y^2)$ links the observed variable y to the hidden value $f(x)$ via a Gaussian noise model. GP assumes that $p(f|X) =$

$p(f(x_1), ..., f(x_N))$ is jointly Gaussian, with mean $m(x) = \mathbb{E}[f(x)]$ and covariance $k(x_i, x_j) = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))^T]$.

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \tag{3.6}$$

Without loss of generality and for the sake of simplicity, the mean function $m(x)$ is commonly set to zero. Function k is a positive definite kernel function, defined based on our prior beliefs over the kinds of functions we expect to observe in data (e.g. level of smoothness). In other words, the kernel function $k(x_i, x_j)$ controls the relativeness of points x_i and x_j, i.e., if the kernel considers x_i and x_j as similar, then output of the function at those points is expected to be similar as well. In this work, we use a radial basis function (RBF) kernel that is in form of $k(x_i, x_j) = \sigma_f^2 \exp(-\frac{1}{2l^2}(x_i - x_j)^2)$. Parameters σ_f and l are optimized based on cross-validation over the training data.

With a new set of unobserved data samples X_*, GP needs to predict f_*. If f_* is to be calculated only based on our prior knowledge (determined by function k), then $f_* \sim \mathcal{N}(0, K(X_*, X_*))$. Figure 3.1-a illustrates 10 sample functions estimated from the GP prior.

GP uses the training samples to train its model and calculate the posterior as follows:

$$\begin{pmatrix} y \\ f_* \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} K_y & K_* \\ K_*^T & K_{**} \end{pmatrix}\right) \tag{3.7}$$

where $K_y = k(X, X) + \sigma_y^2 I_N$ is $N \times N$, $K_* = k(X, X_*)$ is $N \times N_*$, and $K_{**} = k(X_*, X_*)$ is $N_* \times N_*$.

The goal is to compute posterior $p(f_*|X_*, X, y)$ which has the following form:

$$p(f_*|X_*, X, y) = \mathcal{N}(f_*|\mu_*, \Sigma_*) \tag{3.8}$$
$$\mu_* = \mu(X_*) + K_*^T K_y^{-1} y$$
$$\Sigma_* = K_{**} - K_*^T K_y^{-1} K_*$$

derived by applying the rules for conditioning Gaussian distributions. Finally, for each new unobserved sample x_*, GP regressor described above calculates a mean μ_* that is the expected output of the function predicted by GP at the point x, accompanied by a variance σ_*^2 which can be used to interpret the GP's confidence of its prediction.

Figure 3.1-b illustrates 10 sample functions estimated from the GP posterior. As expected, the estimated functions converge to the same output value at the training samples. In Figure 3.1-c, the blue plot is the true function, the dashed red plot is the predicted function by GP (based on μ_*) and the gray areas mark 3 standard deviation away from the predicted function (based on Σ_*).

As a final note, it is good to mention that the Cholesky decomposition is used to compute $K_y^{-1} = L^{-T} L^{-1}$ instead of direct inversion of the matrix since it is faster and also to avoid numerical stability issues (suggested by Rasmussen and Williams [2006]).

3.2.2 Classification

The prediction of labels that Gaussian process provides is probabilistic and the confidence of each prediction can be formally calculated in terms of statistics. This

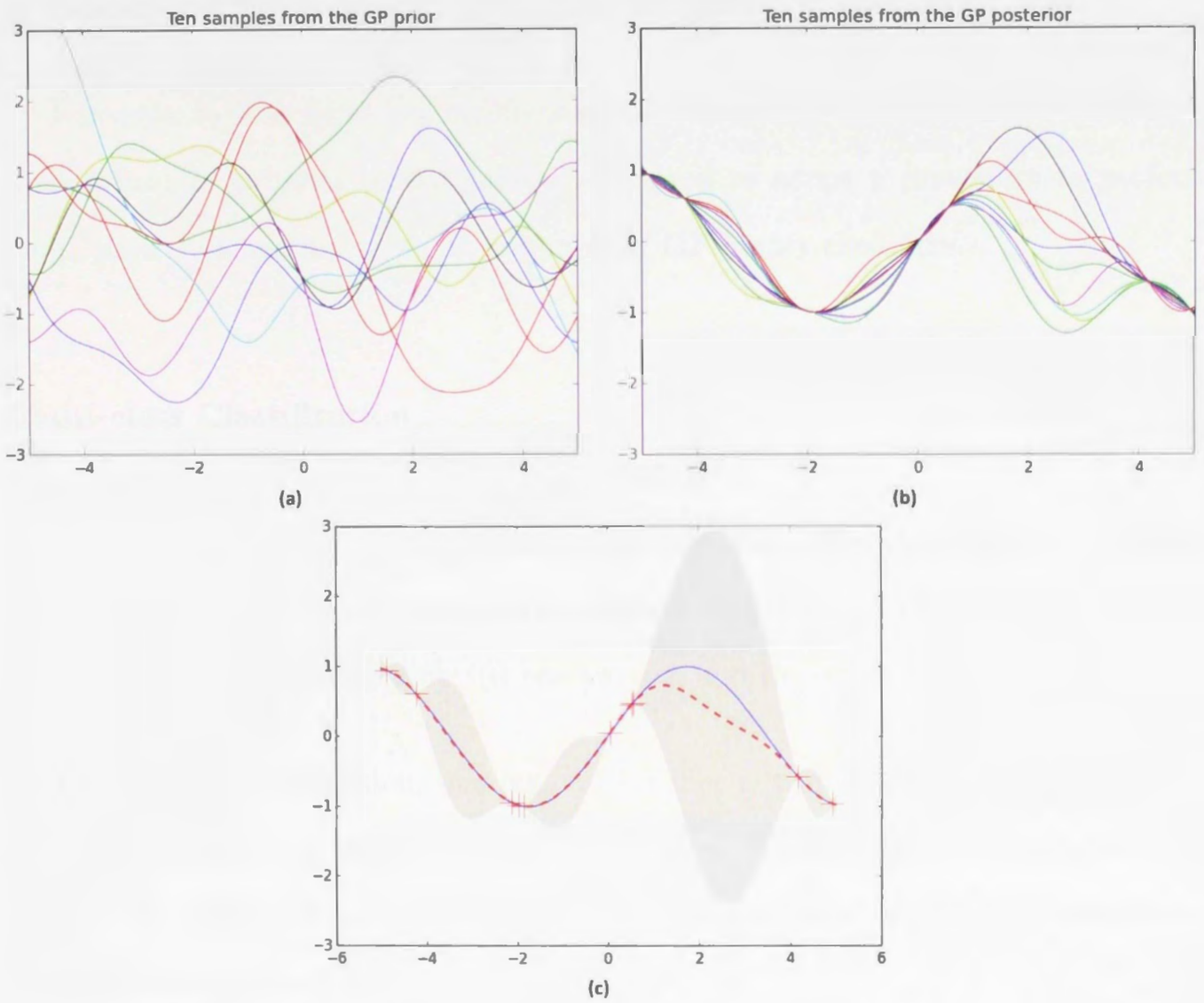

Figure 3.1: (a) 10 sample functions estimated from the GP prior; (b) 10 sample functions estimated from the GP posterior; (c) the blue plot is the true function, the dashed red plot is the predicted function by GP (based on μ_*) and the gray areas mark 3 standard deviation away from the predicted function (based on σ_*)

contrasts with the conventional kernel-based methods such as Support Vector Machines that only provide a guess at the class label which is not associated with a formal confidence estimate.

Following the procedure suggested by Rasmussen and Williams [2006], GPs can be easily converted to binary classifier. To do so, we could assign either +1 or -1 labels to the outputs of the observed data samples, i.e., $y \in \{-1, +1\}$. Then, once the model is trained and the posterior distribution is calculated, for unobserved X_*,

we compute μ_*, where $sign(\mu_*)$ determines the predicted class label.

Since the task at hand is a multi-class classification (i.e., identifying each subject from a host of subjects in the gallery), we need to adopt a procedure to perform multi-class classification using an ensemble of GP binary classifiers.

Multi-class Classification

The binary classification can be extended to cover multi-class classification problems. There are two conventional strategies for reducing the task of multi-class classification to multiple binary classification: (i) one vs. one and (ii) one vs. rest.

In one vs. one reduction, one binary classifier is trained to distinguish between every two classes; this means for a K-class task, $\frac{K(K-1)}{2}$ classifiers are needed to be trained. A voting scheme is then applied on the results of all $\frac{K(K-1)}{2}$ classifiers to come up with the final prediction of the model.

In one vs. rest (a.k.a. one vs. all) reduction, only K classifiers are trained to distinguish between each class versus the rest of classes in data. In this strategy, classifiers should not only be able to predict the class labels, but also provide a confidence score for their decision. In case multiple classifiers predict a positive label for a test data sample, the confidence scores are used for disambiguation by ranking the labels and picking the most probable one.

In the current study, we have addressed our multi-class classification problem with a one vs. rest strategy. The rationale behind selecting this strategy is that much fewer classifiers are needed to be trained, which makes more sense due to the high computational complexity of training a GP-based classifier.

3.3 Face Detection

Since the objective of this study is face identification, it is more convenient and a common practice in the literature to first detect and crop faces from each image and only pass the subjects' faces to the recognizer. Therefore, it is necessary to apply a prior algorithm to track/detect and automatically crop the faces from each video frame. In the following of this section, the two most cited algorithms for face detection that are used in this study are introduced and briefly described.

3.3.1 Viola-Jones method

The face detection algorithm proposed by Viola and Jones [2004] is fast and robust. It consists of three stages:

In the first stage, it uses a representation model called the "Integral Image" that allows for fast computation of the features from images. This representation model includes three kinds of features (see Figure 3.2):

1. The value of a two-rectangle feature is the difference between the sum of the pixels within two rectangular regions.

2. The regions have the same size and shape and are horizontally or vertically adjacent.

3. A three-rectangle feature computes the sum within two outside rectangles subtracted from the sum in a center rectangle. Finally a four-rectangle feature computes the difference between diagonal pairs of rectangles.

There are some limitations associated with the above-mentioned rectangle features, such as being sensitive to the presence of edges, bars, and other simple image structures. However, empirical results show that such representation supports effective learning while being computationally efficient.

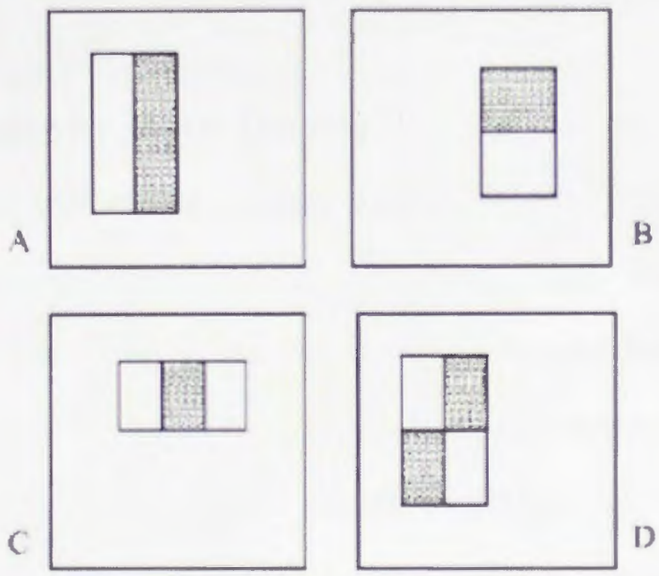

Figure 3.2: Sample rectangle features. Two-rectangle features are shown in (A) and (B). (C) shows a three-rectangle feature. and (D) a four-rectangle feature. Figure is adopted from the paper authored by Viola and Jones [2004].

In the second stage, a simple classifier is trained to select a few critical visual features from a very large set of potential features. The classifier is AdaBoost proposed by Schapire et al. [1998] which in this work, has been used to build a greedy feature selector. AdaBoost aggregates the predictions of a large set of (weak) learners via a weighted majority vote. The weight given to each weak learner determines the importance of that learner, and in this case, the feature.

In the third stage, the classifiers are combined in a "cascade" setting that allows for discarding the background regions of the image so that the focus remains solely on the promising face-like regions. Each layer of the cascade classifier only lets through the sub-windows in the image that it predicts to be a positive one, i.e., partially

containing the face. The final output of the cascade classifier would determine he area of the face in each stationary image.

3.3.2 Incremental learning for Visual Tracking

The Incremental learning for Visual Tracking (IVT) algorithm proposed by Ross et al. [2008] attempts to deal with non-stationary data (videos) where both the target object and the background change over time (i.e., camera motion). This algorithm efficiently learns and updates a low dimensional subspace representation of the target object. The target object is first modelled by a compact representation structure that facilitates object recognition. The subspace model is continuously updated to reflect the changes in appearance of the target object via an efficient incremental method which allows for tracking.

3.4 Feature Extraction

In order to experiment with different feature types, and show that the proposed algorithms would work well with any of them. we use 3 common visual features used in the literature, namely, histogram equalized intensity levels, Local Binary Pattern (LBP) codes, and Histogram of Oriented Gradient (HOG) descriptors. The calculation procedure of of these features are briefly discussed in the rest of this section.

3.4.1 Histogram Equalization

Histogram equalization is a contrast enhancement technique by adjusting the intensities of image pixels by evenly distributing them in the image histogram. This adjustment results in a higher contrast for those regions previously with a lower local contrast. Histogram equalization greatly enhances the quality of the images with poor illumination. This method is specifically useful for low quality videos. Figure 3.3 shows an image before and after histogram equalization.

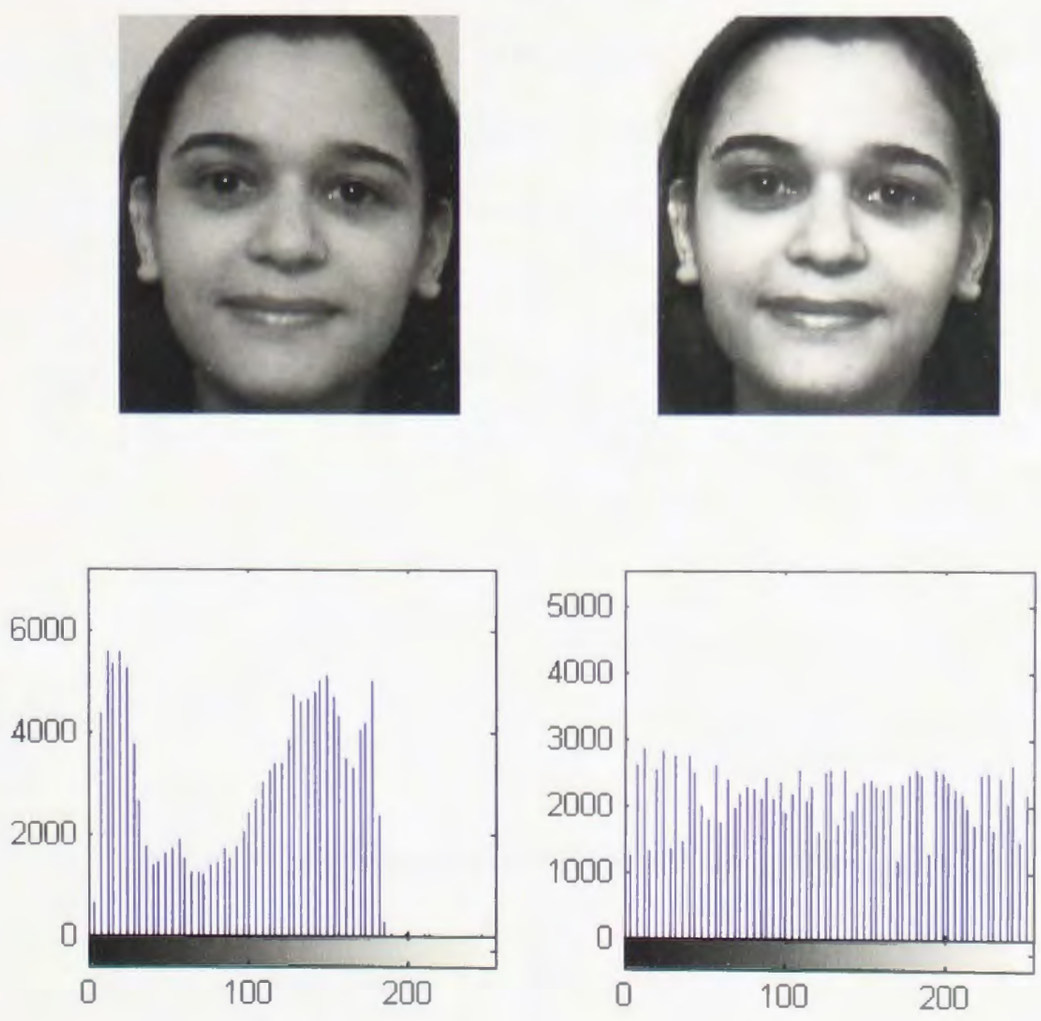

Figure 3.3: Image before (left), and after (right) histogram equalization. The histogram of intensity levels of each image are illustrated in the second row.

3.4.2 Local Binary Patterns

Local Binary Pattern (LBP) operator, first introduced to the machine vision community by Ojala et al. [1996], is a gray-scale invariant texture descriptor. LBP operator is known to be robust to monotonic gray-scale variation which allows for elimination of the effect of poor illumination in images. Also, its computational simplicity makes it a perfect tool for image analysis in challenging real-time settings. This section is an overview of how to derive the LBP codes. A more comprehensive explanation of this process is given in Pietikäinen [2011].

Figure 3.4 illustrates an arbitrary 3×3 block cropped from a gray-level image. The LBP pattern for the pixel in the center (named as g_c) is derived by thresholding the values of its neighbouring pixels (named as g_{pi}) with respect to the value of g_c:

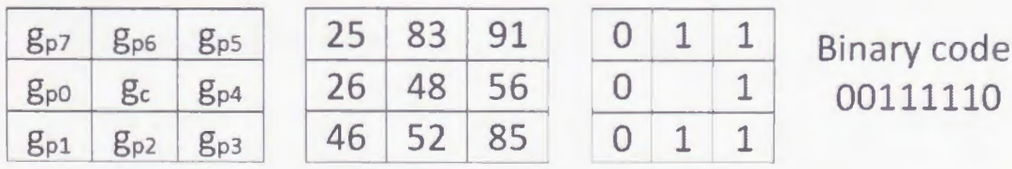

Figure 3.4: Local Binary Pattern operator

$$LBP\ code = \sum_{i=0}^{P-1} sign(g_{pi} - g_c)2^i \tag{3.9}$$

where P is the number of neighbouring pixels and the $sign$ function is defined as:

$$sign(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{3.10}$$

Once the local binary patterns are calculated for every pixel throughout the whole

image/patch, the LBP code for that image/patch is derived as the histogram of these patterns. Please note that for P neighbouring pixels, there will be 2^P different LBP patterns. However, LBP does not use 2^P bins in constructing the histogram. In other words, some patterns would fall in the same bin. To perform the binning, the LBP patterns are classified into two categories: uniform and non-uniform.

The LBP pattern of a pixel is considered uniform if there are none or two transitions between its binary code.[3] It is easy to see that there are $P(P-1)$ patterns that have two transitions.[4] Also, there are 2 patterns that have no transitions.[5] The rest of patterns which are non-uniform fall into the last bin. Therefore, the final histogram (i.e., LBP code) consists of $P(P-1)+3$ bins.

Finally, the histogram generated on each patch of the image is concatenated which forms a vector called the *LBP feature vector*. Please note that in this work, we use the normalized version of this LBP feature vector, meaning that all images construct a unit vector when represented by their respective LBP feature vector.

3.4.3 Histogram of Oriented Gradients

Histogram of Oriented Gradients (HOG) descriptor was first proposed by Dalal and Triggs [2005] for the purpose of object detection in general and human detection in particular. This method is based on the idea that the local appearance of an object can be described by the distribution of intensity gradients (i.e., edge directions).

To calculate the HOG descriptor, each image is first divided into small connected

[3]For instance, the binary code stated in Figure 3.4 contains two transitions: one from the 2^{nd} bit to the 3^{rd}, and the other from the 7^{th} bit to the 8^{th}.

[4]Choose $(P-1)$ different number of 0s (or 1s), then put them in (P) desired places

[5]Either all bits are 0, or all are 1.

blocks, and a histogram of gradient directions is calculated for the pixels within each block. This is derived by filtering each block with a derivative mask such as $[-1, 0, 1]$ (horizontal edge detector), $[-1, 0, 1]^T$ (vertical edge detector), and other more complex masks such as Sobel mask. The HOG descriptor is represented by the concatenation of these histograms (see Figure 3.5).

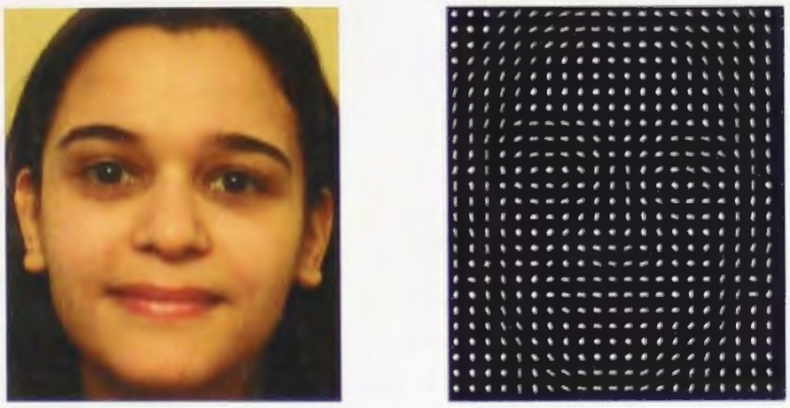

Figure 3.5: Histogram of Oriented Gradients. Left: raw image; Right: extracted HOG features (using blocks of 16×16 pixels)

3.5 Welch t-test

For performance comparison purposes between different methods, we need an statistical tool to compare the accuracies (mean $\pm$ standard deviation) derived by different methods. Welch's t-test (or unequal variances t-test), proposed by Welch [1947], is a two-sample test that is used to test the hypothesis that two populations have equal means.

Welch's t-test is performed when the assumption of equal variance between two populations is not satisfied. In this study, the accuracy of different algorithms vary at difference rate (i.e., difference variances), therefore, the appropriate test is Welch

t-test.

Additionally, Welch t-test is more reliable than the more commonly used student t-test when the two samples have unequal sample sizes. This property comes handy if there is no access to the implemented code for an algorithm so that we can run it with the exact same train/test partition of data, however, we still need to compare our results against the respective method's.

Welch's t-test defines the statistic t by the following formula:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \tag{3.11}$$

where $\bar{X}_1$, s_1^2 and N_1 are the first sample's mean, variance, and size respectively.

The Welch-Satterthwaite equation is used to calculate the degrees of freedom ν associated with this variance estimate:

$$\nu = \frac{(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2})^2}{\frac{s_1^4}{N_1^2 \nu_1} + \frac{s_2^4}{N_2^2 \nu_2}} \tag{3.12}$$

where $\nu_1 = N_1 - 1$ and $\nu_2 = N_2 - 1$ are the degrees of freedom associated with the first and second variance estimates respectively. The approximate degree of freedom is rounded down to the nearest integer.

These two calculated statistics, t and ν, can be used with the t-distribution to test the null hypothesis that the two population have equal means (using a two-tailed test). Alternatively –as in this study– a one-tailed test can be used to evaluate the hypothesis that the mean of one population is greater than or equal to the other.

Chapter 4

Video-based Face Identification

In this chapter, the proposed methods for video-based face identification are discussed.

In Section 4.1, the concepts of quantum formalism are applied to the task of face identification in videos, and the approach to prediction of identity is elaborated in that framework. This approach is referred to as "Quantum Probability Inspired Framework" (QPIF).

In Section 4.2, a dual extension of the quantum framework — referred to as "Ensemble of Abstract Sequence Representatives" (EASR) — is introduced and the identification procedure in this model is discussed.

In section 4.3, a machine learning based approach is added on top of EASR in order to capture the non-linearity available in data and maximize the prediction performance. Gaussian process (GP) is used to implement the machine learning side of the algorithm. GP is a probabilistic but non parametric method that poses no

hard assumption on structure of the model. A novel learning scheme – referred to as Specialization-Generalization – is proposed to support efficient learning for the GP models. Finally, a fast identification process combines predictions of both the GP and EASR modules to efficiently identify the subject in a probe sequence.

4.1 Quantum Probability Inspired Framework

In an image-set based face identification task, each individual image may not fully characterize the face of the person in the image-set. This may be due to (i) poor quality of the image (e.g. low resolution, illumination, etc.), (ii) partial existence of the face in the image's field of view (e.g. occlusion, pose, etc.), and (iii) failure of the face detector algorithm to accurately spot the face. This issue, as mentioned earlier, causes uncertainty on how much a face identification method can rely on each individual image in an image-set (both for known and unknown identities).

Quantum theory can model such uncertainties with its comprehensive notions of probability calculation. It provides a good construct to exploit the knowledge embedded in the image-sets and its representation structure can be employed to solve the complex task of face identification. First, we need to map the concepts of quantum theory in physical world to the problem of face identification with image-sets. Assume there exist a Hilbert space H which includes all the possible images – either observable or unobservable – from all individuals. Each image is represented as a feature vector in the Hilbert space.

4.1.1 Image-sets of known identities as events

Each image-set among the training image-sets (i.e, gallery) belonging to the individual i accounts for a known identity. Each known identity is represented as the corresponding set of feature vectors which span the subspace $S_i \subset H$ (i.e., equivalent to the concept of event in quantum formalism). Each subspace represents one single identity which is equivalent to the notion of *event* or *observable* in the original theory of quantum physical systems. In other words, each subspace S_i is a collective representative of the partial information derived from different images of the individual i. More interestingly, S_i implicitly contains the unobserved images of the individual i as described by the features extracted from the observed images (e.g., a specific texture).

It is also worth mentioning that adding a new subject to the gallery can be easily addressed by defining a new event subspace and extending the events space to cover this new subspace. The rest of the algorithm remains intact.

4.1.2 Image-sets of unknown identities as states

Each unknown identity (i.e., images in the corresponding probe image-set) is represented as an ensemble of states. The feature vector of each image in the probe image-set is considered as an elementary object which defines a *pure* state ϕ. Using (3.1), we then construct several *mixed* states ψ as the superposition of few randomly selected *pure* states in order to maximize the number of attributes that each state can represent, as well as to minimize the effect of uncontrolled variants in the images. Superposition leads to generating initial states that are more robust than the noisy single images and therefore actively improving the recognition process.

In order to represent the unknown identity, it is required to define a probability distribution $p(\psi)$ over the ensemble of participating *mixed* states. The distribution is defined in such a way that each $p(\psi_i)$ reflects a relative inter-similarity measure of the random *pure* states $\{\phi_{j,j=1:n}\}$ that construct the i^{th} *mixed* state ψ_i:

$$p(\psi_i) = \frac{D_i}{\sum_{t=1}^{N} D_t} \quad , \quad D_i = \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} d(\phi_j, \phi_k) \tag{4.1}$$

where metric $d(x, y)$ computes the Euclidean distance between points x and y; n is the number of *pure* states constructing each *mixed* state; and N is the number of *mixed* states in ensemble. Finally, $p(\psi_i)$ is determined by (4.1).

4.1.3 Recognition

Now that the concepts of *event* and *ensemble of states* are clarified in the context of face identification in image-sets, we describe the process of recognition of an unknown identity when the system is presented with a new unseen test image-set. In this section, we explain how QPIF assigns the most probable identity to the test image-set represented by an ensemble of initial states via searching in the events space generated from the training image-sets.

Without loss of generality, let us assume we have M image-sets for the known identities. Then, when presented with an image-set of an unknown identity, we generate N initial states. Now, the task of recognition is carried out as follows: first, the similarity between all initial states $\{\psi_{j,j=1:N}\}$ and events $\{S_{i,i=1:M}\}$ is calculated following (3.4). This results in the *Projection* matrix $P_{M\times N}$ where each element p_{ij}

represents $q(S_i|\psi_j)$:

$$P = \begin{pmatrix} q(S_1|\psi_1) & \dots & q(S_1|\psi_N) \\ \vdots & \ddots & \vdots \\ q(S_M|\psi_1) & \dots & q(S_M|\psi_N) \end{pmatrix} \tag{4.2}$$

where the i^{th} row of matrix P (i.e., $\{q(S_i|\psi_j)_{,j=1:N}\}$) represents the quantum probabilities of each initial state belonging to the known identity S_i. Then, if it is inserted into (3.5). along with the probability distribution of initial states derived by (4.1), the $\{\psi_{j|j=1:N}\}$ will be marginalized out and the probability of the event S_i being true (i.e., the probability that this probe image-set belongs to the known identity represented by S_i) is calculated. However, prior to this process, we have passed the distribution $\{q(S_i|\psi_j)_{,j=1:N}\}$ through a shaping function:

$$q_s(S_i|\psi_j) = q(S_i|\psi_j).exp(1 - \frac{q(S_i|\psi_j)}{\underset{k=1:N}{\text{Max}}\ \{q(S_i|\psi_k)\}}) \tag{4.3}$$

where $exp(x)$ represents the exponential function e^x.

This shaping function was selected empirically based on the experimental results. We will employ the shaped distribution $\{q_s(S_i|\psi_j)_{,j=1:N}\}$ to derive $q(S_i)$. The $q(S_i)$ is the result of marginalizing $\{q_s(S_i|\psi_j)_{,j=1:N}\}$ over the set of all initial states $\{\psi_{j|j=1:N}\}$. The $q(S_i)$ is calculated for all probabilistic events following this procedure and finally, the event with the highest probability is reported as the system's prediction of the identity of the unknown (i.e., probe) image-set. As described above, QPIF employs a simplified interpretation of the quantum formalism to aggregate the information of the known and unknown identities and use the ensemble of initial states for searching in the events space to assign the most probable identity to each probe image-set.

4.2 Ensemble of Abstract Sequence Representatives

Data representation employed in the EASR approach is inspired by the QPIF structure introduced earlier. We extend QPIF's representation method for unknown identities (i.e., ensemble of states) to represent the known identities as well. As a result, both representations of the known and unknown identities can exploit the advantages of sampling and superposition, such as noise relaxation, etc.

As mentioned earlier, in image-set based face identification, each image may not fully characterize the individual's face. This may be due to (i) poor quality of the image (e.g. low resolution, illumination, etc.), (ii) partial existence of the face in the image's field of view (e.g. occlusion, pose, etc.), and (iii) failure of the face detector algorithm to accurately spot the face. Such issues would cast uncertainty on the degree that a face identification method should rely on each individual image in an image-set (for images in both gallery and probe sets).

EASR is a vector-based representation structure for image-sets that addresses the uncertainties mentioned above as follows: it relaxes the noise in the raw data points by transferring them into a higher level representation structure using stratified sampling and superposition. Then, it calculates the similarities between each pair of train and test sequences. Afterwards, the recognition is performed by finding the most similar known EASRs to different generated unknown EASRs candidates, and then aggregating the identification results of all candidates via majority voting.

In the rest of this section, we first explain the representation structure of EASR, and then discuss the method of similarity measurement between each EASR that is

necessary for either performing the identification task, or ranking different candidates in terms of their similarity to the probe image-set.

4.2.1 Representation

A video sequence (top of Figure 4.1) can be represented as a set of normalized n dimensional feature vectors (each referred to as α) extracted from every frame. However, because such primary representation is prone to noise, it is beneficial to transform these vectors into a noise-relaxed secondary representation structure. Using stratified sampling, we draw (with replacement) a set of α vectors from each stratum (i.e., sequence of frames in a video), which are then grouped into several non-overlapping subsets of size m (i.e., m vectors per subset). Then, for each subset, a new feature vector (represented by β) is constructed using (4.4).

$$\beta = \frac{\gamma}{\|\gamma\|} \quad , \quad \gamma = \sum_{i=1}^{m} \alpha_i \tag{4.4}$$

We refer to these new n dimensional unit feature vectors as Abstract Sequence Representatives (ASRs). [1] Superposition leads to constructing more robust samples by minimizing the effect of undesired variations in single noisy images and therefore actively improving the identification accuracy.

For each sequence we construct a set of ASRs of size M, and refer to it as Ensemble of Abstract Sequence Representatives (EASR). Top of Figure 4.1 shows the first 27 frames of a raw sequence labelled J1. In the middle of Figure 4.1 a subset of ASRs forming the J1's EASR is presented. The idea behind introducing ensembles along

[1] Please note that this is the same concept as generating a *mixed* state from a set of *pure* states as in equation 3.1.

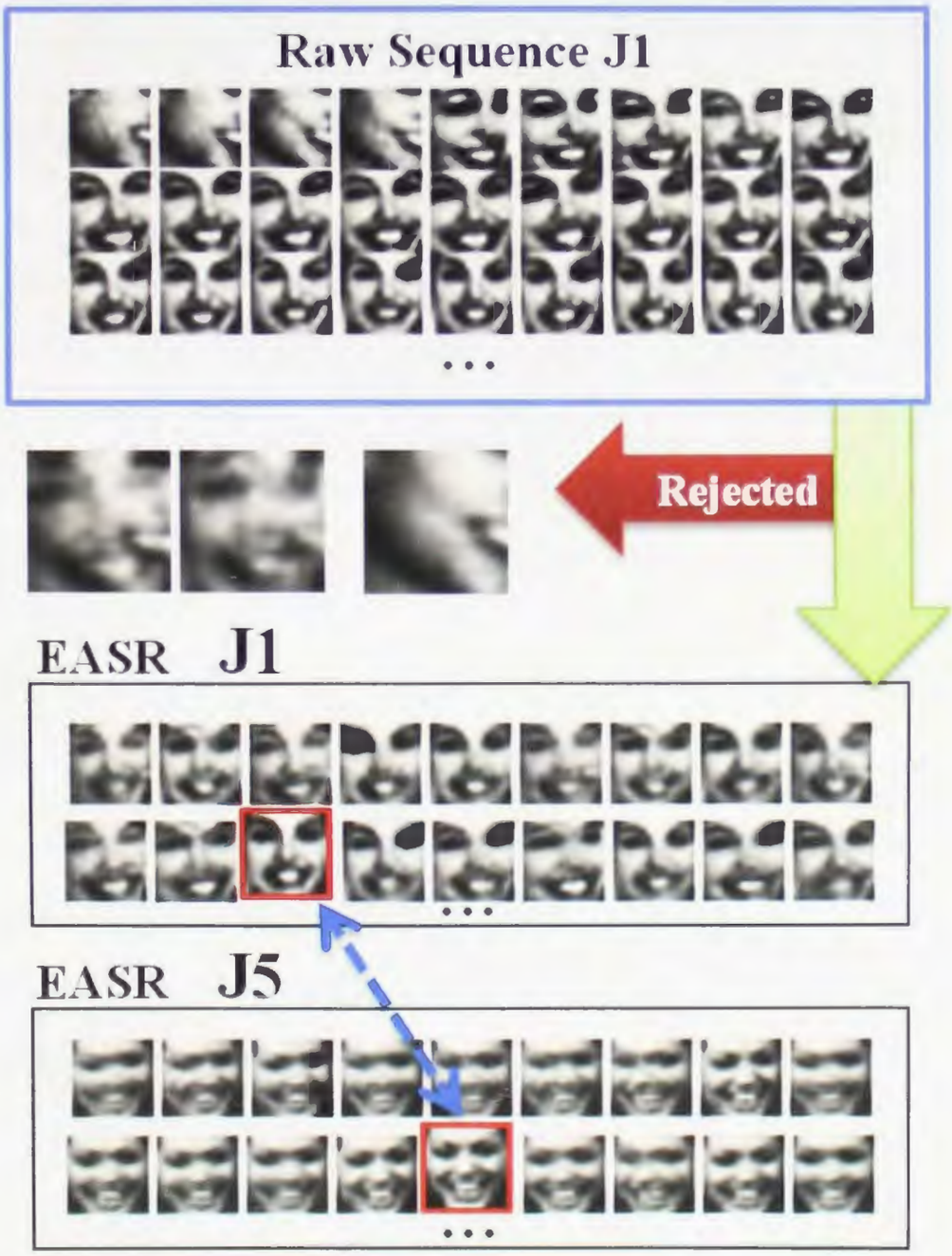

Figure 4.1: A sample sequence from the YouTube Celebrities dataset collected by Kim et al. [2008] (top), its EASR based on intensity features (middle), and a matched EASR from another clip (bottom)

with the majority voting is inspired by the concepts of bagging and exploiting the knowledge of the crowd in decision trees which are known for their robust performance in highly noisy data (see Breiman [1996]).

4.2.2 Similarity Measurement

In order to find the similarity between two video sequences i and j (denoted as S_{ij}), first, we find the similarity ψ_{pq}^{ij} between all possible ASR pairs of the form (β_p^i, β_q^j) where β_p^i is the p^{th} ASR from the EASR of sequence i and β_q^j is the q^{th} ASR from sequence j's ensemble. The pair-wise similarity between two ASRs β_p^i and β_q^j (i.e., ψ_{pq}^{ij}) is calculated as the inner product of the two ASRs, as in (4.5).

$$\psi_{pq}^{ij} = \langle \beta_p^i, \beta_q^j \rangle = |\beta_p^i||\beta_q^j| \cos(\theta_{pq}^{ij}) = \cos(\theta_{pq}^{ij}) \tag{4.5}$$

where θ_{pq}^{ij} is the angle between the two unit vectors β_p^i and β_q^j. All the sequence-wise similarity values ψ_{xy}^{ij} derived by (4.5) are collected in matrix Ψ^{ij} that is an $M \times M$ matrix, where $\Psi^{ij} = [\psi_{pq}^{ij}]$, for $p, q \in [1..M]$. The nearest ASR pair of the two sequences i and j (i.e., the β_p^i and β_q^j with the maximum ψ_{pq}^{ij} among all pairs) determines the similarity measure S_{ij}.

$$S_{ij} = \underset{p=1:M, q=1:M}{\text{Max}} \{\Psi^{ij}\} \tag{4.6}$$

Following our illustrated example, the bottom of Figure 4.1 shows the EASR for a sequence labelled as J5, which represents the same subject as in sequence J1 but belongs to another video clip of hers in the dataset. The similarity between these two EASRs is measured by similarity of their closest pair of ASRs as highlighted in Figure 4.1, calculated via (4.6).

It is a good practice to monitor the quality of ASRs that are generated by random sub-sampling. For each ASR β_p, we calculate its mean pair-wise similarity $\bar{\Psi}_p$ by averaging over row p of Ψ. We then calculate average ($\bar{\Psi}$) and standard deviation ($\sigma_{\bar{\Psi}}$) of these $\bar{\Psi}_p$s for $p \in [1..M]$. Finally, we filter out the possible outliers, i.e., any ASR β_o with an average pair-wise similarity ($\bar{\Psi}_o$) that is two standard deviations ($\sigma_{\bar{\Psi}}$) less than the average within-ensemble similarity ($\bar{\Psi}$).

$$\text{Reject}(\beta_o | \bar{\Psi}_o < \bar{\Psi} - 2\sigma_{\bar{\Psi}}) \tag{4.7}$$

We identified two sources for generating outlier ASRs: (i) superposition of frames that present the subject in highly different conditions; and (ii) presence of noisy frames (e.g., where the face tracking failed) in the ASR. Three sample ASRs that were rejected in the process of constructing the EASR for the J1 sequence are shown in Figure 4.1. The two rejected ASRs on the left are generated due to source (i), while source (ii) is behind rejection of the third ASR (mostly a result of face tracker failure on a number of frames).

4.2.3 Identification

In the identification stage, the votes of several independent decision-makers (i.e., several ensembles of ASRs for each probe video sequence) are aggregated to come up with a robust identity recognition of a sequence belonging to an unknown identity. For each probe video sequence (let us say i^{th}), the proposed method builds several (let us say K) ensembles and calculates their similarity ($S_{i_{kj}}$, $k = 1 : K$) to every ensemble of all training sequences ($j = 1 : Q$). This will generate the matrix $\Lambda_{i_{K \times Q}}$

for the i^{th} test sequence. Therefore, the k^{th} row in Λ_i compares the k^{th} ensemble generated from the test sequence against all the training sequences. Then, for each row of Λ_i, we find the one training sequence that has the maximum similarity to this probe ensemble and report the identity associated with this sequence using (4.8).

$$I_{K\times 1} = index(\underset{row-wise}{\text{ArgMax}}\ \{\Lambda_i\}) \tag{4.8}$$

where $index(x)$ returns the identity associated with the x^{th} image-set.

Afterwards, the function *vote* counts the number of times that each identity has been selected as the most similar one to the ensemble of current probe video sequence and records it in the respective cell in matrix V which consists of P cells, where P is the number of known identities (i.e., number of subjects). [2]

$$V_{P\times 1} = vote(I_{K\times 1}) \tag{4.9}$$

Ultimately, the known identity associated with the highest number of votes is reported as the system's final prediction of identity, following (4.10).

$$identity = \text{ArgMax}\ \{V_{P\times 1}\} \tag{4.10}$$

In case of a tie in number of votes between two or more candidates, a score is assigned to each of them for ranking purpose. This score is calculated for candidate t, denoted as $score_t$, as the sum of similarity values from the ensembles that voted for

[2] $P = Q$ only if there is exactly one training video sequence per subject available in the gallery.

this candidate using (4.11).

$$score_t = \sum_{k \in find(I==t)} \text{Max } \{\Lambda_{i_k}\} \tag{4.11}$$

where the function $find(X == a)$ returns the indices of the elements in vector X which are equal to the scalar a; and Λ_{i_k} is the k$^{\text{th}}$ row of the matrix Λ_i in (4.8). The identity with the highest score is reported as the final prediction of the system.

4.3 Ensemble of Gaussian Process Models on top of EASR, a Hierarchical Approach

In the previous section, Ensemble of Abstract Sequence Representatives (EASR) was introduced that is a vector-based method for representing a video sequence. We mentioned that each EASR is built by sampling and superposition to reduce noise, followed by a filtering mechanism to deal with outliers. EASR models the variations of the subject in an image-set and the similarity of EASRs can be used for identification purposes. However, EASR representation is linear and would fail to capture the non-linear underlying structure of the data.

In order to address the non-linearity in data, we propose to use an ensemble of binary Gaussian process (GP) models in a one-versus-rest setting on top of the EASR representation method. The results is a hierarchy of two main modules: EASR module and GP module. EASR module offers better resistance to noisy data and the GP module incorporates a learning scheme called specialization – generalization for effective training of an ensemble of binary GP classifiers (enabling further noise

reduction). The identification process combines both modules using a hierarchical structure to maximize identification rate. The rest of this section describes the GP module in details.

4.3.1 Gaussian Process Models

In the current work, we use the GP regressor to construct a GP binary classifier (i.e., $y_i \in \{-1, +1\}$). For each subject i, this classifier is capable of identifying the subject i versus the rest of the subjects. For the sake of implementation, let us assume subject i has m_i samples in total for training. We label these samples as $(+1)$. In order to collect the (-1) labelled samples, we sub-sample an equal number of data points from the training set belonging to the rest of the subjects in the gallery such that the total number of samples for training is closest possible to $2m_i$. This is to make sure that the training data samples are balanced and avoid bias in the classifier's decision making process.

For each subject i in a gallery with L subjects, we construct one GP model GP^i. This model is trained to predict whether a sequence belongs to the subject i or not. To predict the identity of a probe video sequence p with m_p frames, the frames are presented to all L models. Each GP^i predicts the μ_* that is a vector of m_p length, where the j^{th} item shows the expected value of GP^i's underlying function (f_*) with the j^{th} frame as input. Classification of each frame is based on the sign of μ_*, if it is negative, it means GP^i rejects the possibility that this frame belongs to the subject i and vice versa. In order to aggregate the outputs of all m_p frames, we calculate the average of all f_*^j, $j \in [1..L]$ and record it as the overall output of GP^i (i.e., sum-fusion). After calculating the aggregated output for every model, identity of the

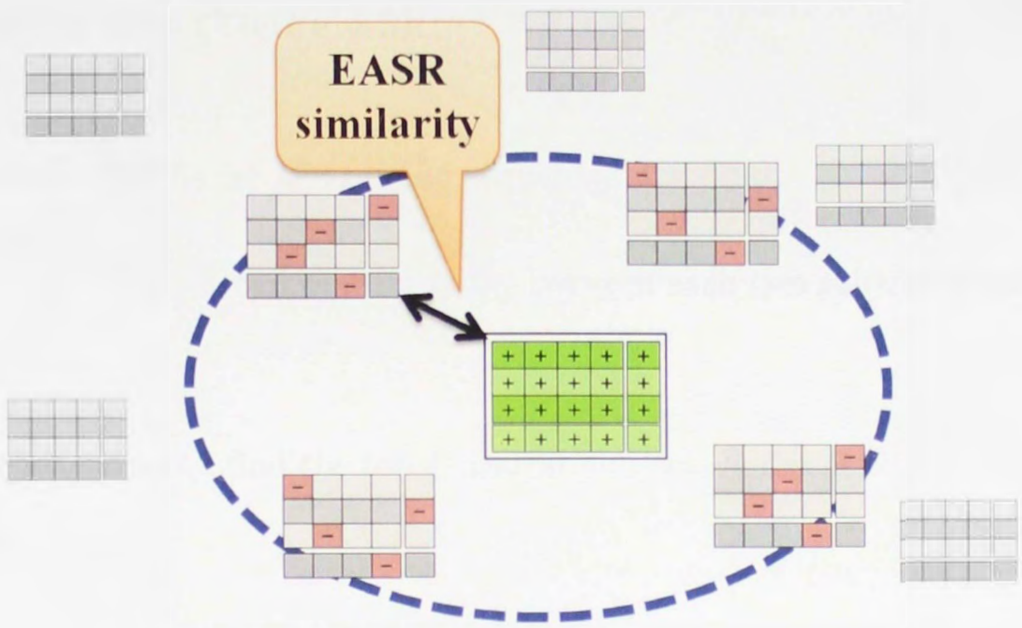

Figure 4.2: The specialization step for $k = 4$ (best viewed in color)

subject with the highest aggregated output is reported as the predicted identity by the GP ensemble.

4.3.2 Specialization – Generalization Learning Scheme

GP binary classifiers are sensitive to the quality of training samples, thus a simple random sampling process without any provision for avoiding noisy samples reduces the identification power of the resulting model. In this section, we describe our learning scheme which relies on EASRs for finding the most relevant sequences for training each binary GP model (i.e., specialization step, schematically shown in Figure 4.2), complemented by a generalization step which tries to alleviate the effect of potentially noisy frames in the training samples.

Starting with n subjects and m sequences for each subject in the training data, we have a $SQ_{n \times m}$ which contains sequences for each subject.

Specialization step (Figure 4.2):

1. Calculate EASRs for all training sequences.
2. Calculate the pair-wise similarity S_{ij} between each two subjects i and j, following (4.6).
3. For each subject i find the top k nearest subjects with the highest S_{ij} and store js in NS_i.
4. Train GP_i with all frames from $SQ_i, [1..m]$ as $(+1)$ instances and randomly sub-sample equal number of frames from $SQ_{j,[1..m]}, j \in NS_i$ as (-1) instances.

Generalization step:

5. Use GP_i to label each sequence in $SQ_{j,[1..m]}, j \notin NS_i$, for each frame f if $\text{GP}_i(f) > 0$ (i.e., mislabelled) add it to $GenL_i$ list to be retrained to GP_i.
6. Update GP_i with all frames f in $GenL_i$ as (-1) instances.

In the specialization step, for each GP model, frames of the training sequences for the target identity are used as $(+1)$ instances. The (-1) instances are randomly sub-sampled from the sequences belonging to the k nearest subjects to the target identity, as determined by the EASR similarity (Figure 4.2). The goal of the specialization step is to force GP to learn distinctive features that separate each subject from the most similar subjects to him/her in the gallery.

However, we have to make sure that the GP binary classifier would generalize well on the subjects whom it has not seen during its first batch of training (i.e.,

Figure 4.3: Sample noisy frames detected in the generalization step when training a GP model for the sequence J1 in Figure 4.1

specialization step). In generalization step, we randomly sub-sample some frames from videos featuring those subjects not used in the specialization step, and evaluate the GP binary classifier. If the label is not correct[3] (the correct label would be (-1) since we know this sample is definitely not featuring the respective subject), then the model is re-trained with this sample as a (-1) instance. In other words, the generalization step provides more (-1) instances to the GP model in areas of the problem space that the model is unable to correctly identify such instances, thus improving the generalizability of the GP model.

The generalization step also minimizes the effect of noisy frames in the $(+1)$ instances. For example, consider the first 3 frames of sequence J1 shown at the top of Figure 4.1. These frames do not provide any useful information for identifying the subject in that video. In the initial training of the GP model for sequence J1, these frames are provided as $(+1)$ instances, which misleads the model to classify any similar noisy frame as $(+1)$. In the generalization step, such noisy frames belonging to the rest of the training sequences are detected. Figure 4.3 shows a selection of these detected frames when training the model for J1. These frames are then used as new (-1) instances to update the GP model. This process helps to successfully cancel out the effect of noisy frames in the $(+1)$ instances.

Now that the GP models have been constructed consulting the EASR suggestions, the next stage is to make predictions based on the models.

[3]Based on the empirical results, the label is correct in most of the samples.

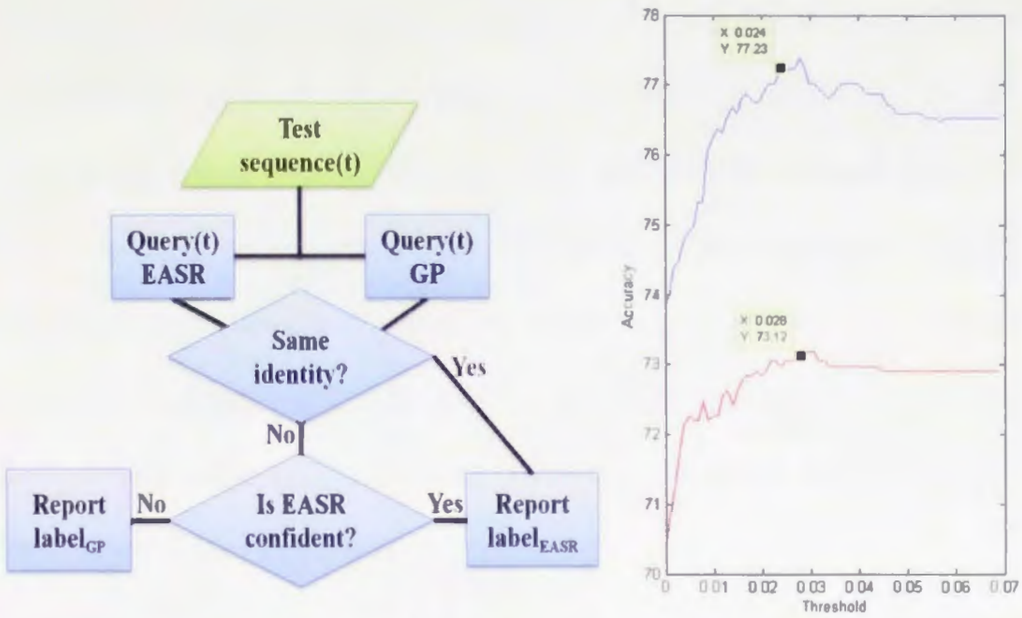

Figure 4.4: Flowchart for the identification process (left); Exploring the effect of minimum cut-off for EASR confidence (τ) on accuracy (right)

4.3.3 Identification: a Hierarchical Approach

In this section, we discuss our proposed hierarchical approach for aggregating the predictions of the two modules, namely EASR module and GP module, to come up with the most accurate prediction of identity for a probe video sequence. Figure 4.4-left illustrates the flowchart of the proposed approach.

Clearly, if predictions of both modules agree on the same identity, that prediction is reported. Otherwise, the hierarchical approach is used as follows: First, we give priority to the EASR module since it is more noise tolerant. We trust EASR-based prediction when it identifies the probe video sequence p by a clear winner; that is, when difference in similarity of p and the winner S_{pw} versus p and the closest next candidate S_{pc} as derived by (4.6) is above a pre-defined threshold τ. If the constraint for the cut-off is not satisfied, it indicates that the EASR module is not confident in its prediction, therefore the label generated by the GP module is reported as the final predicted identity.

Figure 4.4-right shows the overall accuracy of our method for different values of τ in two different experiments (described later). The right side of Figure 4.4 shows that accuracy drops when we rely too much on the EASR module (low τ values). On the other hand, relying too much on the GP module has the same effect. The value of τ for each dataset, is selected based on cross-validation over the training set (the selected points are highlighted in the graphs in Figure 4.4-right). As similarity of two EASRs falls between zero and one, $max(\tau) = 1$. However, in our experiments, τ is much smaller (always less than 0.05).

Figure 5.1: Sample faces extracted from Honda/UCSD dataset

5.1.1 Honda/UCSD

Honda/UCSD dataset is a collection of 59 videos recorded from 20 subjects in order to form a common ground for assessment of different face identification algorithms (see Figure 5.1). Each subject has at least 2 videos (except for one subject). All the videos have an equal resolution of 640×480 and recorded at 15fps rate. Duration of the videos vary from 71 to 645 frames. with mean and standard deviation of 258.2 and 110.8 respectively.

5.1.2 CMU-MoBo

CMU-MoBo dataset was primarily collected for automatic identification of people by gait. However, it has been recently used for image-set based face identification studies as well (see Figure 5.2). This dataset contains video sequences from 6 camera views of 25 subjects performing four different walking activities on a treadmill: slow, fast, on inclined surface, and holding a ball. Following the literature, the subject with fewer than four walking patterns is excluded from the dataset, thus only the first 24 subjects

Chapter 5

Experimental Design

In this chapter, we briefly describe the datasets and evaluation settings for our experiments.

5.1 Datasets

Three publicly available benchmark datasets are used for evaluation of the proposed methods in this study: Honda/UCSD dataset collected by Lee et al. [2003], CMU MoBo dataset collected by Gross and Shi [2001], and the more challenging YouTube Celebrities collected by Kim et al. [2008].

Figure 5.1: Sample faces extracted from Honda/UCSD dataset

5.1.1 Honda/UCSD

Honda/UCSD dataset is a collection of 59 videos recorded from 20 subjects in order to form a common ground for assessment of different face identification algorithms (see Figure 5.1). Each subject has at least 2 videos (except for one subject). All the videos have an equal resolution of 640×480 and recorded at 15fps rate. Duration of the videos vary from 71 to 645 frames, with mean and standard deviation of 258.2 and 110.8 respectively.

5.1.2 CMU-MoBo

CMU-MoBo dataset was primarily collected for automatic identification of people by gait. However, it has been recently used for image-set based face identification studies as well (see Figure 5.2). This dataset contains video sequences from 6 camera views of 25 subjects performing four different walking activities on a treadmill: slow, fast, on inclined surface, and holding a ball. Following the literature, the subject with fewer than four walking patterns is excluded from the dataset, thus only the first 24 subjects

Figure 5.2: Sample faces extracted from CMU-MoBo dataset, provided by Cevikalp and Triggs [2010]

are used. All videos are of 640 × 480 resolution and recorded at 30fps rate. Duration of the videos vary from 202 to 897 frames, with mean and standard deviation of 495.6 and 169.3 respectively.

5.1.3 YouTube Celebrities

YouTube Celebrities dataset is a collection of real-world videos from YouTube website featuring 47 celebrities (see Figure 5.3). The videos are noisy, low resolution, and demonstrate large variations in illumination, pose, expression, and other uncontrolled conditions. For each subject there are 3 video clips, where each clip is divided into several sequences of unequal resolution and duration (between 7 to 350 frames, with mean and standard deviation of 163.0 and 84.5 respectively). There is a total number of 1910 sequences, all encoded in MPEG4 at 25fps rate.

Figure 5.3: Sample faces extracted from YouTube Celebrities dataset; (a), (b), and (c) illustrate samples from 3 different clips of the same person

5.2 Evaluation Settings

In this section, we describe the procedure for preparation of the training and test data. We followed the common settings used in the literature to allow for fair comparison.

5.2.1 Face Detection

It is a common practice to first track and crop faces from each frame and only pass the subjects' faces to the recognizer. Since the objective of this study is face identification, it is more convenient to only pass the subjects' faces to the recognizer. Therefore, it is necessary to apply a prior algorithm to track/detect and automatically crop the faces from each video frame. Similar to the previous works in the literature, Viola-Jones method, proposed by Viola and Jones [2004], is used for extracting faces in the

Honda/UCSD[1] and CMU-MoBo[2] datasets. For the YouTube Celebrities dataset, the Viola-Jones algorithm fails to detect faces in a number of sequences. Thus, following Hu et al. [2012] we [3] use the Incremental learning for Visual Tracking (IVT) algorithm, proposed by Ross et al. [2008]. IVT returns the face area in all frames of all 1910 sequences, however, some may not represent a correct face (see Figure 4.3). [4]

5.2.2 Resolution

All the cropped faces are resized to an equal resolution. Images in Honda/UCSD dataset are resized to 20×20 pixels, CMU-MoBo to 40×40 pixels, and YouTube Celebrities dataset to 20×20 pixels (20×20 resolution was selected to reduce the computational cost).

5.2.3 Features

In order to experiment with different feature types, we use histogram equalized intensity levels for the Honda/UCSD dataset, Local Binary Pattern (LBP) codes, proposed by Ojala et al. [2002], for the CMU-MoBo dataset, and Histogram of Oriented Gradient (HOG) descriptors, proposed by Dalal and Triggs [2005], for the YouTube

[1]The author would like to thank Dr. Liang Chen for providing the Honda/UCSD dataset with faces detected.

[2]In this work, we have directly used the pre-processed version of CMU-MoBo dataset provided by the authors of Cevikalp and Triggs [2010]. The pre-processing procedure includes face tracking, resolution, and feature extraction.

[3]The author would like to thank Dr. Liang Chen for providing the YouTube Celebrities dataset with faces detected using the IVT algorithm.

[4]Wang et al. [2012b] have created another version of this dataset in which they also use the Viola-Jones algorithm, however, the sequences containing frames with falsely detected faces were manually removed. Thus resulting in a dataset with fewer than 1910 video sequences as in the original version of the dataset. In this study, no evaluations were run on the Wang et al. [2012b] version of the YouTube Celebrities dataset.

Celebrities dataset [5].

5.2.4 Train/Test image-set arrangement

For the Honda/UCSD dataset we randomly select 20 sequences (one video per subject) for training and the rest for testing. It should be noted that, there is an alternative evaluation setting for the Honda/UCSD dataset, which uses a *predefined* set of 20 sequences (one video per subject) for training without any random permutations. Since recent algorithms (e.g., RNP, ISCRC, and teh proposed methods) achieve 100% accuracy with this predefined setting, we use the random setting which provides more variation in order to have a more meaningful comparison.

For the CMU-MoBo dataset we also randomly select 24 sequences, one video per subject for training and the rest for testing.

For the YouTube Celebrities dataset we perform 5-fold cross-validation, following the evaluation protocol used by Hu et al. [2012]. Sequences of each subject are sequentially partitioned (no prior shuffling) into 5 folds, where each fold contains exactly 9 sequences (from 3 clips) with minimal overlap between folds. In each fold, 1 clip is randomly selected for training (3 sequences) and the other 2 clips are used as test data (6 sequences).

It is important to mention that there is another evaluation setting for the YouTube Celebrities dataset first used by Wang et al. [2012b]. In this setting, for every subject in each fold, 9 sequences (3 per clip) is randomly selected; 3 sequences (1 per clip) for training, and the rest for testing.

[5]The author would like to thank Dr. Liang Chen for providing his code for efficient calculation of the HOG features.

Obviously, it is an easier task to identify the subject with the second setting, because there is already one video sequence from each clip available in the training set, which factors out differences in appearance of the subject in different clips. For this reason, we believe that the first setting is closer to real world scenarios, thus we adopted the protocol used by Hu et al. 2012

For all three datasets we report accuracy results for the full length sequences as well as truncated sequences that only contain the first 50 consecutive frames of each video sequence. All evaluations are done using 5-fold cross validation

Chapter 6

Results and Discussions

In this section, we summarize the identification accuracies of the three proposed approaches and compare them against the most successful and recent methods in the literature (namely, MSM, MDA, AHISD/CHISD, SANP, RNP, MSSRC, JSR, and IS-CRC in chronological order). Except for JSR, for all other methods we used the code provided by the authors adjusted with their suggested parameter values. For JSR we did not have access to the code thus report the results provided by the authors. However, it should be noted that the evaluation settings for JSR are different than what we are using in this study – they used the Wang et al. setting for the YouTube Celebrities dataset (which leads to higher accuracy results compared to the Hu et al. [2012]'s settings used here), and 30×30 resolution for both YouTube Celebrities and CMU-MoBo datasets.

To make the comparisons fair, we used the same evaluation settings, including feature type for training all algorithms (i.e., intensity levels for Honda/UCSD, LBP for CMU-MoBo, and HOG for YouTube Celebrities). Interestingly, this enhancement

led to improved accuracy for all algorithms (including the older algorithms such as SANP) on the YouTube Celebrities dataset compared to the results reported in the original papers. Also, it must be noted that the original evaluation of RNP was done only on 29 subjects for the YouTube Celebrities dataset, and the results obtained in the respective paper (Yang et al. [2013]) are higher than the results obtained on the full dataset. Additionally, MSSRC (Ortiz et al. [2013]) comes with its own face tracking algorithm which was disabled in our evaluations, since the aim is to compare only the identification power of different algorithms, therefore, the same tracking algorithm is used for all evaluations.

Performance results on each of the three benchmark datasets is derived by exactly following the protocol described in the Evaluation Settings section in chapter 5. This protocol is the same as that in the related works in the literature to allow for fair comparison. We perform Welch's t-test (see Welch [1947]) to check whether the improvement in performance of the proposed method(s) is statistically significant compared to the best performance of the contender methods. Outcomes of the significance tests are described along with the summary of performance results.

Tables 6.1-6.3 summarize the *Mean* $\pm$ *Standard Deviation* of the identification rates for different methods in the literature and the three methods proposed in this work (namely, QPIF, EASR, and EASR+GP) on Honda/UCSD, CMU-MoBo, and YouTube Celebrities datasets for both the truncated sequences (only the first 50 frames are available to perform the identification task), as well as the full length video sequences.

On the Honda/UCSD dataset, both QPIF and EASR methods performed well for the full length video sequences and EASR+GP outperformed al other methods for both truncated as well as the full length video sequences. The difference in identifi-

Table 6.1: Identification Rates (%) of Different Methods on Honda/UCSD Dataset (Mean ± Standard Deviation)

Method	**Year**	50 frames	Full length
MSM	1998	87.69 ± 6.12	90.26 ± 2.15
MDA	2009	87.69 ± 2.81	96.41 ± 1.40
AHISD	2010	88.21 ± 3.89	83.59 ± 3.89
CHISD	2010	86.15 ± 2.92	91.28 ± 2.29
SANP	2011	87.18 ± 6.01	96.41 ± 2.29
RNP	2013	88.21 ± 4.66	93.33 ± 2.29
MSSRC	2013	91.28 ± 1.40	93.85 ± 2.29
ISCRC	2014	92.31 ± 4.44	95.38 ± 1.15
QPIF		91.28 ± 2.29	$\mathbf{97.44 \pm 3.63}$
EASR		91.28 ± 1.40	$\mathbf{98.46 \pm 1.40}^{\diamond}$
EASR+GP		$\mathbf{94.87 \pm 4.05}$	$\mathbf{99.49 \pm 1.15}^{\star}$

Note:

$\star$ indicates statistically significant improvement of accuracy compared to the second best result at $\alpha = 0.05$

$\diamond$ indicates statistically significant improvement of accuracy compared to the second best result at $\alpha = 0.1$

cation rates for EASR+GP method is significantly (statistical) higher than the best contending method in the literature.

On the CMU-MoBo dataset (see Table 6.2), the EASR+GP method achieved a slightly lower identification rate compared to ISCRC (less than 1%). However, based on the statistical analysis, there is no significant difference between the results.

It should be noted that the Honda/UCSD and CMU-MoBo datasets are commonly used as benchmarks and considered as easier recognition tasks since most of the algorithms in the literature have already achieved above 90% accuracy. Therefore, there

Table 6.2: Identification Rates (%) of Different Methods on CMU-MoBo Dataset (Mean ± Standard Deviation)

Method	**Year**	50 frames	Full length
MSM	1998	92.50 ± 2.71	97.22 ± 1.70
MDA	2009	84.17 ± 6.56	95.28 ± 2.88
AHISD	2010	92.50 ± 2.71	95.56 ± 2.48
CHISD	2010	92.50 ± 2.71	98.61 ± 1.39
SANP	2011	92.50 ± 2.71	99.17 ± 0.76
RNP	2013	92.50 ± 2.71	98.33 ± 1.16
MSSRC	2013	91.11 ± 3.49	98.33 ± 1.52
ISCRC	2014	$\mathbf{94.44 \pm 2.20}^{\dagger}$	$\mathbf{99.44 \pm 0.76}^{\dagger}$
QPIF		92.50 ± 2.71	97.50 ± 1.81
EASR		91.94 ± 3.32	$98.89 \pm 1.16^{\dagger}$
EASR+GP		$93.61 \pm 2.71^{\dagger}$	$98.89 \pm 1.16^{\dagger}$

Note: † indicates no significant difference between the best performance result (bold) and the proposed approaches (statistically)

is not much room for improvement. However, we believe that the results on the most challenging dataset, YouTube Celebrities, can rank different algorithms in terms of performance and efficiency.

For the YouTube Celebrities dataset (see Table 6.3), QPIF and EASR slightly outperformed the contending methods on the full length video sequences and the EASR+GP approach achieved significantly better results and improved state-of-the-art by $\approx$ 4% for the full length sequences. EASR+GP also achieves the highest accuracy for the truncated sequences. The superior results of the proposed method can be attributed to its capability of handling extremely noisy samples in the YouTube Celebrities dataset more efficiently compared to the rest of the methods in the literature.

Table 6.3: Identification Rates (%) of Different Methods on YouTube Celebrities Dataset (Mean ± Standard Deviation)

Method	**Year**	50 frames	Full length
MSM	1998	70.57 ± 5.33	65.82 ± 4.56
MDA	2009	64.26 ± 3.76	69.22 ± 4.90
AHISD	2010	69.43 ± 4.16	63.83 ± 3.24
CHISD	2010	67.73 ± 5.09	69.65 ± 4.59
SANP	2011	67.59 ± 5.71	73.40 ± 3.18
RNP	2013	69.50 ± 5.30	73.48 ± 3.65
MSSRC	2013	70.78 ± 3.48	72.20 ± 3.52
ISCRC	2014	66.38 ± 4.73	70.71 ± 3.14
QPIF		69.01 ± 4.36	$\mathbf{74.82 \pm 2.49}$
EASR		70.43 ± 4.16	$\mathbf{74.18 \pm 3.35}$
EASR+GP		$\mathbf{73.12 \pm 3.11}$	$\mathbf{77.23 \pm 3.81}^{\diamond}$

Note: $\diamond$ indicates statistically significant improvement of accuracy compared to the second best result at $\alpha = 0.1$

It is also worth mentioning that a simple ensemble of GP binary classifiers without employing the specialization – generalization learning strategy performed poorly on YouTube Celebrities, testifying to the merit of our approach.

As mentioned before, the competing methods also benefited from using HOG features, especially the top two performers, namely RNP and SANP. When HOG features are used, the average accuracy of RNP and SANP algorithms increase by over 8% compared to the reported accuracies in the respective papers (Yang et al. [2013]) and (Hu et al. [2012]) in which the intensity levels were used as features.

The identification rate for JSR on the YouTube Celebrities dataset with full length video sequences is 73.7% as reported in work by Cui et al. [2014]. This accuracy is only 0.2% higher than the best contender reported here (and it does not affect the

result of the significance test). However, as mentioned before, the evaluation setting used in Cui et al. [2014] is different and the results are reported for 30×30 resolution, therefore we did not include this result in Table 6.3.

QPIF and EASR methods performed very close to each other in all three datasets, while as expected, EASR+GP performed better than both of these methods. The combination of GP and EASR enabled us to achieve a better performance, noticeably higher than the individual components of the method (EASR and GP). We attribute this increase in identification rate to the EASR's strength in dealing with noisy frames, as well as the GP's strength in capturing underlying non-linear structures in data.

6.1 Computational Complexity

We also report the average computation time of all methods in experiments on the YouTube Celebrities dataset for the truncated sequences (with 50 frames). All the timing results are reported based on running Matlab codes provided by the authors of each algorithm on a machine with an Intel Xeon E5-2603 (1.8 GHz) processor and 40 Gigabytes of RAM. We report the average online identification time (in seconds) for one sequence (Table 6.4). We also provide the total offline training time (in seconds) for methods that required training including the EASR+GP method.

While our proposed method requires an initial offline training of the models (over 70% of this time is used for training the GP models), it is important to note that the offline training time is a one-time only overhead. For comparison, SANP will require an extra 160 seconds for identifying only 10 test sequences compared to our method. Also, adding a new subject to the gallery requires far less training time, since only

Table 6.4: Average computation time (seconds) of different methods on the YouTube Celebrities dataset with truncated sequences (50 frames). T1: Total offline training time. T2: Average online testing time for one sequence.

	MSM	MDA	AHISD	CHISD	SANP	RNP	MSSRC	ISCRC	EASR+GP
T1	N/A	24.21	N/A	N/A	N/A	1.18	N/A	22.31	156.18
T2	0.64	2.63	3.35	8.46	19.01	0.92	70.78	2.04	2.79

Note: N/A indicates online-only methods

one new model needs to be constructed.

Chapter 7

Conclusions and Future Work

7.1 Summary of Contributions

- Introduction of two non-sophisticated representation structures using the notions of quantum probability theory, namely, QPIF and its dual extension EASR. The proposed representation structures are designed to minimize the effect of noisy frames in a video based face identification task. These two representation structures specifically target those frames that are not useful for the identification task, mainly due to face occlusion, low resolution of image, or failure of the face tracker algorithm. Therefore, unlike most of the methods in the literature which use sophisticated non-linear representations, these two methods keep the representation linear and simple while retaining the superior performance.

- A novel learning scheme was proposed for efficient training of an ensemble of binary Gaussian process models. This learning scheme selectively samples from

the training data in order to not only increase the discrimination power of the classifier, but also to build its models using the least possible computational cost and with minimum introduction of noise.

As a final note, the contributions of this work are not method-specific and can be utilized for enhancement of other face identification approaches in the literature.

7.2 Future Work

- In the current work, each representative is either accepted or rejected to contribute in building the model and predicting the identity. A promising extension of this work would be to modify the outlier filtering process in EASR by utilizing a probabilistic approach that is able to assign a degree of uncertainty on how well each frame is a good representative of an individual's face.

  This should be accompanied by a prediction method that can exploit the extra information provided by such weighted samples. Consequently, we will have a classification approach that is aware of the quality of the samples and knows on which of them it should rely the most in order to perform the identification task effectively.

- Employ EASR approach along with its outlier filtering process as a general purpose filtering approach to improve other methods in the literature in terms of their resilience to noisy frames.

- Use other kernels (e.g., pyramid match kernel proposed by Grauman and Darrell [2007]) for the Gaussian process models which can better take advantage of

localized feature descriptors such as Scale-Invariant Feature Transform (SIFT) proposed by Lowe [2004].

- Go beyond the face identification task and test the proposed method with datasets for other types of video based recognition tasks (e.g., object categorization).

Bibliography

O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 581–588, 2005.

Leslie E Ballentine. The statistical interpretation of quantum mechanics. *Reviews of Modern Physics*, 42(4):358, 1970.

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

H. Cevikalp and B. Triggs. Face recognition based on image sets. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2567–2573, 2010.

Yi-Chen Chen, VishalM. Patel, P.Jonathon Phillips, and Rama Chellappa. Dictionary-based face recognition from video. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision, ECCV 2012*, volume 7577 of *Lecture Notes in Computer Science*, pages 766–779. Springer Berlin Heidelberg, 2012.

Zhen Cui, Hong Chang, Shiguang Shan, Bingpeng Ma, and Xilin Chen. Joint sparse representation for video-based face recognition. *Neurocomputing*, 135(0):306 – 312,

2014. ISSN 0925-2312. doi: http://dx.doi.org/10.1016/j.neucom.2013.12.004. URL http://www.sciencedirect.com/science/article/pii/S092523121301148X.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

Carl Henrik Ek, Philip HS Torr, and Neil D Lawrence. Gaussian process latent variable models for human pose estimation. In *Machine learning for multimodal interaction*, pages 132–143. Springer, 2008.

Kazuhiro Fukui and Osamu Yamaguchi. Face recognition using multi-viewpoint patterns for robot vision. In *Robotics Research*, pages 192–201. Springer, 2005.

Kristen Grauman and Trevor Darrell. The pyramid match kernel: Efficient learning with sets of features. *The Journal of Machine Learning Research*, 8:725–760, 2007.

Ralph Gross and Jianbo Shi. The cmu motion of body (mobo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Pittsburgh, PA, June 2001.

Lov K Grover. A fast quantum mechanical algorithm for database search. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 212–219. ACM, 1996.

Negar Hassanpour and Liang Chen. A hierarchical training and identification method using gaussian process models for face recognition in videos. In *The 11th IEEE International Conference on Automatic Face and Gesture Recognition*, 2015a.

Negar Hassanpour and Liang Chen. A quantum theory inspired framework for face identification in videos. Technical report, University of Northern British Columbia, Department of Computer Science, Computational Intelligence Laboratory, 2015b.

Yiqun Hu, Ajmal S Mian, and Robyn Owens. Face recognition using sparse approximated nearest points between image sets. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(10):1992–2004, 2012.

Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. Active learning with gaussian processes for object categorization. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

Kihwan Kim, Dongryeol Lee, and Irfan Essa. Gaussian process regression flow for analysis of motion trajectories. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1164–1171. IEEE, 2011.

Minyoung Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

Tae-Kyun Kim, Josef Kittler, and Roberto Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):1005–1018, 2007.

Kuang-Chih Lee, J. Ho, Ming-Hsuan Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–313–I–320 vol.1, 2003.

David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture

learning for robust visual tracking. *International Journal of Computer Vision*, 77 (1-3):125–141, 2008. ISSN 0920-5691.

Robert E Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of statistics*, pages 1651–1686, 1998.

Paul Viola and Michael Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57:137–154, 2004.

John von Neumann and Robert T Beyer. *Mathematical foundations of quantum mechanics*. Princeton University Press, 1955.

Ruiping Wang and Xilin Chen. Manifold discriminant analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 429–436, 2009.

Ruiping Wang, Huimin Guo, L.S. Davis, and Qionghai Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2496–2503, June 2012a. doi: 10.1109/CVPR.2012.6247965.

Ruiping Wang, Shiguang Shan, Xilin Chen, Qionghai Dai, and Wen Gao. Manifold-manifold distance and its application to face recognition with image sets. *IEEE Transactions on Image Processing*, 21(10):4466–4479, 2012b.

Tiesheng Wang and Pengfei Shi. Kernel grassmannian distances and discriminant analysis for face recognition from image sets. *Pattern Recognition Letters*, 30(13): 1161–1165, 2009.

Bernard L Welch. The generalization of ŝtudent'spŕoblem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.

O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. In *Proceedings of 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, pages 318–323, 1998.

Meng Yang, Pengfei Zhu, L. Van Gool, and Lei Zhang. Face recognition based on regularized nearest points between image sets. In *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pages 1–7, 2013.

Pengfei Zhu, Wangmeng Zuo, Lei Zhang, S.C.-K. Shiu, and D. Zhang. Image set-based collaborative representation for face recognition. *Information Forensics and Security, IEEE Transactions on*, 9(7):1120–1132, 2014.